

ASYMPTOTIC NORMALITY OF MAXIMUM LIKELIHOOD AND ITS VARIATIONAL APPROXIMATION FOR STOCHASTIC BLOCKMODELS

BY PETER BICKEL , DAVID CHOI , XIANGYU CHANG , AND HAI ZHANG

University of California, Berkeley

University of California, Berkeley

Xi'an Jiaotong University

Northwest University

Variational methods for parameter estimation are an active research area, potentially offering computationally tractable heuristics with theoretical performance bounds. We build on recent work that applies such methods to network data, and establish asymptotic normality rates for parameter estimates of stochastic blockmodel data, by either maximum likelihood or variational estimation. The result also applies to various sub-models of the stochastic blockmodel found in the literature.

1. Introduction. The analysis of network data is an open statistical problem, with many potential applications in the social sciences [13] and in biology [15]. In such applications, the models tend to pose both computational and statistical challenges, in that neither their fitting method nor their large sample properties are well-understood.

However, some results are becoming known for a model known as the stochastic blockmodel, which assumes that the network connections are explainable by a latent discrete class variable associated with each node. For this model, consistency has been shown for profile likelihood maximization [1], a spectral-clustering based method [16], and other methods as well [2, 6, 7, 8], under varying assumptions on the sparsity of the network and the number of classes. These results suggest that the model has reasonable statistical properties, and empirical experiments suggest that efficient approximate methods may suffice to find the parameter estimates. However, formally there is no satisfactory inference theory for the behavior of classical procedures such as maximum likelihood under the model, nor for any procedure which is computationally not potentially NP under worst-case analysis.

AMS 2000 subject classifications: network statistics, stochastic blockmodeling, variational methods, maximum likelihood

In this note, we establish both consistency and asymptotic normality of maximum likelihood estimation, and also of a variational approximation method, considering sparse models and restricted sub-models under similar assumptions as required in [1]. To some extent, we are following a pioneering paper of Celisse et. al. [5], in which the dense model was considered, and consistency was established for a subset of the parameters.

2. Preliminaries.

2.1. General graph models. We consider a class of latent variable models for unlabeled graphs considered by various authors[11, 1, 4], which we describe as follows. Let Z_1, \dots, Z_n be \mathcal{Z} -valued latent random variables, corresponding to vertices $1, \dots, n$, let π be a distribution on \mathcal{Z} , and let h be a symmetric map $\mathcal{Z} \times \mathcal{Z} \rightarrow [0, 1]$. We define the *complete graph model* (CGM) for (\underline{Z}, A) , where $\underline{Z} = (Z_1, \dots, Z_n)$ and A is the $n \times n$ symmetric 0-1 adjacency matrix of a graph, by its density with respect to an appropriate reference measure.

$$(1) \quad f(\underline{Z}, A) = \left(\prod_{i=1}^n \pi(Z_i) \right) \left(\prod_{i=1}^n \prod_{j=i+1}^n h(Z_i, Z_j)^{A_{ij}} (1 - h(Z_i, Z_j))^{1-A_{ij}} \right),$$

where we may interpret $h(Z_i, Z_j)$ as $\mathbb{P}(\text{edge} | Z_i, Z_j)$.

The *graph model* (GM) is defined by distribution $g : \{0, 1\}^{n \times n} \rightarrow [0, 1]$, which satisfies $g(A) = \mathbb{P}(A; h, \pi)$, in addition to the identity

$$(2) \quad \frac{g(A)}{g_0(A)} = \mathbb{E}_{f_0} \left[\frac{f(\underline{Z}, A)}{f_0(\underline{Z}, A)} \mid A \right],$$

which holds for any g_0 and f_0 corresponding to the same choice of h and π .

It is data from GM which we assume we observe. In [1], $\{Z_i\}_{i=1}^n$ are i.i.d Uniform $(0, 1)$. In [10], they are a multivariate mixture of Gaussians with unknown parameters. If we make no restrictions on h , these models are equivalent.

2.2. Stochastic Blockmodel. In this paper, we focus on stochastic blockmodels, in which \mathcal{Z} is a discrete space $\{1, \dots, K\}$. As a result, we may simplify the GM model by letting π be a discrete distribution over $\{1, \dots, K\}$, and by letting h be represented by a matrix $H \in [0, 1]^{K \times K}$.

We will consider parametric submodels of the blockmodel. Parameterized CGM and GM densities f and g for the blockmodel generically satisfy

$$f(\underline{Z}, A; \theta) = \left(\prod_{i=1}^n \pi_\theta(Z_i) \right) \left(\prod_{i=1}^n \prod_{j=i+1}^n H_\theta(Z_i, Z_j)^{A_{ij}} (1 - H_\theta(Z_i, Z_j))^{1-A_{ij}} \right),$$

and

$$g(A; \theta) = \sum_{\underline{Z} \in \mathcal{Z}^n} f(\underline{Z}, A; \theta).$$

We will sometimes use a parameterization $\theta = (\rho, \phi)$ in which

$$\begin{aligned} H_\theta &\equiv \rho S_\phi \\ \pi_\theta &\equiv \pi_\phi, \end{aligned}$$

where $\rho > 0$ is a nonnegative scalar; ϕ is a Euclidean parameter ranging over an open set; S_ϕ is a symmetric matrix in $\mathbb{R}^{K \times K}$ constrained to satisfy $\sum_{a,b=1}^K \pi_\phi(a) \pi_\phi(b) S_\phi(a, b) = 1$; and the map $\phi \mapsto (\pi_\phi, S_\phi)$ is assumed to be smooth. Let $\lambda = n\rho$. The interpretation of these parameters is that $\lambda = \mathbb{E}[\text{degree}]$, $\rho = \mathbb{P}(A_{ij} = 1)$, and

$$\pi(a) \pi(b) S(a, b) = \mathbb{P}(Z_i = a, Z_j = b | A_{ij} = 1).$$

We will use this parameterization to analyze asymptotic behavior when $\rho \equiv \rho_n \rightarrow 0$ while keeping ϕ is fixed, as seems reasonable for sparse network settings.

An interesting class of submodels, discussed by Newman [12], are the “degree-corrected” blockmodels with UV classes obtained by considering $Z_i = (Z_{i1}, Z_{i2})$, for $i = 1, \dots, n$, which take values (u, v) ; where u takes values $1, \dots, U$ with probabilities $\alpha_1, \dots, \alpha_U$; and given parameters $\gamma_1, \dots, \gamma_V \in [0, 1]$, v takes values $\gamma_1, \dots, \gamma_V$ with probabilities β_1, \dots, β_V . We will assume Z_{i1} and Z_{i2} are independent. Additional parameters needed are a $U \times U$ symmetric matrix of probabilities G . We can now define

$$\mathbb{P}(Z_{i1} = a, Z_{i2} = \gamma_c, Z_{j1} = b, Z_{j2} = \gamma_d | A_{ij} = 1) = \alpha_a \alpha_b \beta_c \beta_d \gamma_c \gamma_d G(a, b).$$

So although this is a UV blockmodel, it has only $U(U+1)/2 + (U-1) + (2V-1)$ parameters. Its interpretation is that there are U subblocks, but within each subblock, vertices can hierarchically exhibit further affinities to vertices both within the same block and other blocks, thus enabling, for instance, distinction between vertices of high degree and low degree within each block. This distinction is not block-dependent, resulting in a reduction of parameters.

Many variants are of course possible; for example, one can choose to have more parameters by having the (u, v) block probabilities be free, so that the conditional distribution of Z_{i2} dependent on Z_{i1} , or fewer parameters by treating $\alpha(1), \dots, \alpha(U)$ as known.

2.3. *Maximum likelihood and variational estimates.* For the complete graph blockmodel, maximum likelihood estimation of H and π (or of θ) is basically understood. From Eq. (1) it can be seen that the log likelihood expression decomposes, so that π is estimated from \underline{Z} independently of A , and H is estimated from A conditional on \underline{Z} . We note that it is possible for the likelihood to have multiple local optima; in particular this is the case for the degree-corrected blockmodel CGM.

For the GM blockmodel, the maximum likelihood parameter estimate $\hat{\theta}^{ML}$ is given by

$$\begin{aligned}\hat{\theta}^{ML} &= \arg \max_{\theta} g(A; \theta) \\ &= \arg \max_{\theta} \sum_{\underline{Z} \in \mathcal{Z}^n} f(\underline{Z}, A, \theta).\end{aligned}$$

Multiple local optima in g may exist even if the CGM likelihood function f is concave in the appropriate parameterization, as we shall see for the ordinary unrestricted parameterization. Additionally, the maximum likelihood estimate involves a generally intractable marginalization over the latent variable \underline{Z} .

Variational methods attempt to circumvent the second difficulty (while accepting the first) by introducing an approximate function J for which local optimization is computationally easier. For the GM blockmodel, the estimate $\hat{\theta}^{VAR}$ is given by

$$\begin{aligned}\hat{\theta}^{VAR} &= \arg \max_{\theta} \max_{q \in \mathcal{D}} J(q, \theta; A) \\ &\triangleq \arg \max_{\theta} \max_{q \in \mathcal{D}} -D(q || f_{\underline{Z}|A;\theta}) + \log g(A; \theta).\end{aligned}$$

Here \mathcal{D} is the set of all product distributions over \mathcal{Z}^n , with densities denoted by $\prod_{i=1}^n q_i(\cdot)$. The term $D(\cdot || \cdot)$ is the Kullback-Leibler divergence, and $f_{\underline{Z}|A;\theta}$ is the conditional density of \underline{Z} given A , i.e., $f_{\underline{Z}|A;\theta}(Z) = \frac{f(Z, A; \theta)}{g(A; \theta)}$. The Kullback-Leibler divergence is given by

$$D(q || f_{\underline{Z}|A;\theta}) = \sum_{Z \in \mathcal{Z}^n} q(Z) \log \frac{q(Z)}{f_{\underline{Z}|A;\theta}(Z)}.$$

We note that J simplifies to

$$\begin{aligned}J(q, \theta; A) &= \sum_{i=1}^n \sum_{a=1}^K q_i(a) [-\log q_i(a) + \log \pi_{\theta}(a)] \\ &\quad + \sum_{i=1}^n \sum_{j=i+1}^n \sum_{a=1}^K \sum_{b=1}^K q_i(a) q_j(b) [A_{ij} \log H_{\theta}(a, b) + (1 - A_{ij}) \log(1 - H_{\theta}(a, b))].\end{aligned}$$

This formula indicates that, at least for the complete parameterization, a local optimum to J can be tractably computed for moderate n and K using the EM algorithm as in [9]. In contrast, optimization of g requires a summation over \mathcal{Z}^n which is generally intractable. However, note that we have added $n(K - 1)$ new parameters.

We remark that $\exp(J) \leq g$ always, due to the nonnegativity of the Kullback-Leibler divergence. Intuitively, we expect the variational estimate to approximate the maximum likelihood estimate when there exists $q \in \mathcal{D}$ which is close to $f_{Z|A;\theta}$. We also remark that J takes a similar form under the general model with continuous \mathcal{Z} and parameters (π, h) , suggesting that the approximation may have utility in that setting as well.

3. Results.

3.1. Asymptotic normality of maximum likelihood under CGM blockmodel.

We first review the asymptotics of the CGM block model with complete parameterization.

Let $\varpi \in \mathbb{R}^K$ and $\mu \in \mathbb{R}^{K \times K}$ be the logit of the π and H , given by

$$\begin{aligned} \varpi(a) &= \log \frac{\pi(a)}{1 - \sum_{b=1}^{K-1} \pi(b)} & a = 1, \dots, K-1 \\ \mu(a, b) &= \log \frac{H(a, b)}{1 - H(a, b)} & a, b = 1, \dots, K. \end{aligned}$$

Given data \underline{Z}, A generated by the model, let f_0 and ϖ_0, μ_0 correspond to the generative parameter values. For the CGM blockmodel, the log likelihood ratio $\Lambda = \log \frac{f}{f_0}$ as a function of $\theta \equiv (\varpi, \mu)$ is given by

$$\begin{aligned} \Lambda(\theta, \underline{Z}, A) &= \sum_{a=1}^{K-1} \left[(\varpi(a) - \varpi_0(a))n_a - n \log \frac{1 + \sum_{a=1}^{K-1} e^{\varpi(a)}}{1 + \sum_{a=1}^{K-1} e^{\varpi_0(a)}} \right] \\ &\quad + \sum_{a=1}^K \sum_{b=1}^K \left[(\mu(a, b) - \mu_0(a, b))O_{ab} - n_{ab} \log \frac{1 + e^{\mu(a, b)}}{1 + e^{\mu_0(a, b)}} \right], \end{aligned}$$

where

$$\begin{aligned} O_{ab} &= \sum_{i=1}^n \sum_{j=i+1}^n 1\{Z_i = a, Z_j = b\} A_{ij}, & n_a &= \sum_{i=1}^n 1\{Z_i = a\}, \\ n_{ab} &= \sum_{i=1}^n \sum_{j=i+1}^n 1\{Z_i = a, Z_j = b\}. \end{aligned}$$

This is an exponential family in θ with gradient $\nabla\Lambda$ given by

$$\begin{aligned} (\nabla\Lambda)_{\varpi(a)} &= n_a - n\pi(a) & a = 1, \dots, K-1 \\ (\nabla\Lambda)_{\mu(a,b)} &= O_{ab} - n_{ab}H(a,b) & a = 1, \dots, K, \ b = a, \dots, K, \end{aligned}$$

provided standard regularity conditions, e.g., θ is an interior point of the canonical parameter space, and $\mathbb{E}(\nabla\Lambda)(\nabla\Lambda)^T$ is of rank $K(K+3)/2 - 1$ uniformly on the set $\{\theta : |\theta - \theta_0| \leq M\}$.

LEMMA 1 (Local asymptotic normality). *For the CGM with parameter values ϖ_0, μ_0 , under standard regularity conditions it holds for any s, t that*

$$\Lambda\left(\varpi_0 + \frac{s}{\sqrt{n}}, \mu_0 + \frac{t}{\sqrt{n^2\rho_0}}\right) = \Lambda(\varpi, \mu) + sY_1 + tY_2 - \frac{1}{2}s^T\Sigma_1s - \frac{1}{2}t^T\Sigma_2t + o_P(1),$$

where $Y_1 \sim N(0, \Sigma_1)$ and $Y_2 \sim N(0, \Sigma_2)$ and are independent, and Σ_1 and Σ_2 are functions of ϖ_0 and μ_0 .

PROOF. Taylor expand and apply central limit theorem. \square

LEMMA 2. *Assume that $n\lambda \rightarrow \infty$, that $0 < \pi_a < 1$ for $a = 1, \dots, K$, and that $S_{ab} > 0$ for all a, b . Then if \hat{S}^{CGM} and $\hat{\pi}^{CGM} = (\hat{\pi}(1), \dots, \hat{\pi}(K-1))$ are maximum likelihood estimates for the CGM with parameters π_0, S_0 , it holds under standard regularity conditions that under (π_0, S_0) ,*

$$\begin{aligned} \sqrt{n}(\hat{\pi}^{CGM} - \pi_0) &\rightarrow N(0, \Sigma_1) \\ \sqrt{n\lambda}(\hat{S}^{CGM} - S_0) &\rightarrow N(0, \Sigma_2). \end{aligned}$$

PROOF. Standard exponential family theory, e.g. [3]. \square

3.2. *Asymptotic normality of maximum likelihood under GM blockmodel.* Our main result is that if (\mathcal{Z}, A) are data from a blockmodel with generative parameter $\theta_0 \in \mathbb{R}^p$, and $\hat{\theta}^{ML}$ achieves the local optima of $\theta \mapsto g(A, \theta)$ closest to θ_0 , then asymptotic normality of $\hat{\theta}^{ML}$ holds.

DEFINITION 1. *A classification of \mathcal{Z} is any onto mapping $\hat{Z} : \{1, \dots, n\} \rightarrow \{1, \dots, K\}$ which depends only on A .*

An essentially correct classification to order γ_n on $\bar{\Theta} \subset \Theta$ is one such that for all $\theta_0 \in \bar{\Theta}$,

$$(3) \quad \sup_{\theta \in \bar{\Theta}} \mathbb{P}_\theta(\mathcal{Z} \neq \hat{Z}(A) | A) \frac{g(A; \theta)}{g(A; \theta_0)} = o_P(\gamma_n)$$

$$(4) \quad \sup_{\theta \in \bar{\Theta}} \mathbb{P}_\theta(\mathcal{Z}, A : \mathcal{Z} \neq \hat{Z}(A)) = o(\gamma_n).$$

We will use a result from [1] to establish that an essentially correct classification exists for the blockmodel under certain conditions. We defer proof of Theorem 1 to a later section.

THEOREM 1. *Let \underline{Z}, A be generated from a blockmodel, parameterizable as $\theta \equiv (\rho, \pi_\phi, S_\phi)$, such that ϕ is fixed, S_ϕ has no identical columns, and $\rho = \rho_n$ satisfies $n\rho_n/\log n \rightarrow \infty$. Let $\bar{\Theta} = \left(\frac{\log n}{n}, 1\right) \times \Phi$, where Φ is an open compact set. For all $c > 0$, there exists an essentially correct classification to order $\gamma_n(K) = o(n^{-c})$ on $\bar{\Theta}$.*

THEOREM 2. *Suppose an essentially correct classification to order $\gamma_n(K) = o(1)$ on $\bar{\Theta}$ holds, and that (\underline{Z}, A) are generated from the model with parameter $\theta_0 \in \bar{\Theta}$. Then for all $\theta \in \bar{\Theta}$,*

$$(5) \quad \frac{g}{g_0}(A, \theta) = \frac{f}{f_0}(\underline{Z}, A, \theta)(1 + \epsilon_n(K, \theta)) + \epsilon_n(K, \theta),$$

where $\sup_{\theta \in \bar{\Theta}} \epsilon_n(K, \theta) = o_P(1)$.

PROOF OF THEOREM 2. Let 1_E to denote the indicator on the event $E = \{\underline{Z}, A : \underline{Z} = \hat{Z}(A)\}$.

$$\begin{aligned} \frac{g}{g_0}(A, \theta) &= \mathbb{E}_{\theta_0} \left[\frac{f}{f_0}(Z, A, \theta) | A \right] \\ &= \mathbb{E}_{\theta_0} \left[\frac{f}{f_0}(Z, A, \theta) 1_E | A \right] + \mathbb{E}_{\theta_0} \left[\frac{f}{f_0}(Z, A, \theta) 1_{\bar{E}} | A \right] \\ &= \frac{f}{f_0}(\hat{Z}, A, \theta) \mathbb{P}_{\theta_0}(E|A) + \mathbb{P}_{\theta_0}(\bar{E}|A) \frac{g}{g_0}(A, \theta) \\ &= \frac{f}{f_0}(\underline{Z}, A, \theta) \mathbb{P}_{\theta_0}(E|A) \\ &\quad + \left(\frac{f}{f_0}(\hat{Z}, A, \theta) - \frac{f}{f_0}(\underline{Z}, A, \theta) \right) \mathbb{P}_{\theta_0}(E|A) 1_E + \mathbb{P}_{\theta_0}(\bar{E}|A) \frac{g}{g_0}(A, \theta) \\ &= \frac{f}{f_0}(\underline{Z}, A, \theta)(1 - o_P(1)) + o_P(1) + o_P(1) \end{aligned}$$

where in the last equation we used $\mathbb{P}_{\theta_0}(E|A) = 1 - o_P(1)$ by Eq. (3) (noting that $\frac{g}{g_0}(A, \theta_0) = 1$); used $\left(\frac{f}{f_0}(\hat{Z}, A, \theta) - \frac{f}{f_0}(\underline{Z}, A, \theta)\right) \mathbb{P}_{\theta_0}(E|A) 1_E = o_P(1)$ (since $\mathbb{P}(E) = o(1)$ by Eq. (4)); and used $\mathbb{P}_{\theta_0}(\bar{E}|A) \frac{g}{g_0}(A, \theta) = o_P(1)$ by Eq. (3). The $o_P(1)$ terms are uniform over all $\theta \in \bar{\Theta}$. \square

THEOREM 3. Assuming the conditions of Theorem 1 and 2, let $\hat{\pi}^{ML}, \hat{S}^{ML}$ and $\hat{\pi}^{CGM}, \hat{S}^{CGM}$ be the corresponding maximum likelihood estimates over all $\theta \in \bar{\Theta}$. It holds that

$$(6) \quad \begin{aligned} \hat{\pi}^{ML} - \hat{\pi}^{CGM} &= o_P(n^{-1/2}) \\ \hat{S}^{ML} - \hat{S}^{CGM} &= o_P((n\lambda)^{-1/2}). \end{aligned}$$

PROOF OF THEOREM 3. By Theorems 1 and 2, the maximizers and local maximizers of the CGM and GM likelihoods must be consistent since $\frac{f}{f_0}(Z, A, \theta_0) = \frac{g}{g_0}(A, \theta_0) = 1$, so that $\epsilon(K, \theta)$ is uniformly negligible. It follows that $\left| \frac{f}{f_0}(Z, A, \theta^{CGM}) - \frac{f}{f_0}(Z, A, \theta^{ML}) \right| = o_P(1)$. The conclusion follows by Lemma 1, which states that $\log \frac{f}{f_0}$ has nonvanishing curvature at θ_0 . \square

We thus have established asymptotic normality not only for the block-model under the full parameterization, but also for submodels such as the degree-corrected variant. For a general parameterization, we have $\hat{\phi}^{ML} - \hat{\phi}^{CGM} = o_P(n^{-1/2})$ generically. If ϕ is separable into (ϕ_π, ϕ_S) such that $\pi = \pi_{\phi_\pi}$ and $S = S_{\phi_S}$, and ϕ_π and ϕ_S are allowed to vary freely, then $\hat{\phi}_S^{ML} - \phi_S$ is asymptotically normal with the faster rate $\sqrt{n\lambda}$, assuming standard regularity conditions. Independence of $\hat{\phi}_S^{ML}$ and $\hat{\phi}_\pi^{ML}$ is then also valid as well. Also, the same holds for the CGM case.

3.3. *Asymptotic normality of variational estimates under GM blockmodel.* The same properties that we have established for maximum likelihood estimates under the GM blockmodel also hold for the more computable variational likelihood estimates.

THEOREM 4. Let $J(\theta; A)$ denote $\max_{q \in \mathcal{D}} \exp[J(q, \theta; A)]$. Under the conditions of Theorem 1 and 2,

$$(7) \quad \frac{J(\theta; A)}{g(A; \theta_0)} = \frac{f}{f_0}(\tilde{Z}, A, \theta)(1 + o_P(1)) + o_P(1),$$

and hence the conclusions of Theorem 3, apply to π^{VAR}, S^{VAR} , the variational likelihood estimates.

PROOF. As usual, we let g_0 and θ_0 correspond to the generative model. We establish Eq. (7).

Let $\delta_{\hat{Z}}$ denote the indicator function $\delta_{\hat{Z}}(Z) = 1\{Z = \hat{Z}(A)\}$. We observe that

$$\exp[J(\delta_{\hat{Z}}, \theta; A)] = f(\hat{Z}(A), A; \theta),$$

and it thus follows that

$$\begin{aligned}
 \frac{f}{f_0}(\hat{Z}(A), A; \theta) &= \frac{\exp[J(\delta_{\hat{Z}}, \theta; A)]}{f(\hat{Z}, A; \theta_0)} \\
 &\leq \max_{q \in \mathcal{D}} \frac{\exp[J(q, \theta; A)]}{f(\hat{Z}, A; \theta_0)} \\
 &= \max_{q \in \mathcal{D}} \frac{\exp[J(q, \theta; A)]}{g(A; \theta_0) f(\hat{Z}, A; \theta_0) g(A; \theta_0)^{-1}} \\
 &= \max_{q \in \mathcal{D}} \frac{\exp[J(q, \theta; A)]}{g(A; \theta_0) \mathbb{P}_{\theta_0}(\mathcal{Z} = \hat{Z}(A) | A)} \\
 &= \max_{q \in \mathcal{D}} \frac{\exp[J(q, \theta; A)]}{g(A; \theta_0) (1 - o_P(1))},
 \end{aligned}$$

where the $o_P(1)$ term converges uniformly over all $\theta \in \bar{\Theta}$. Rearranging terms, it therefore follows that

$$\begin{aligned}
 \max_{q \in \mathcal{D}} \frac{\exp[J(q, \theta; A)]}{g(A; \theta_0)} &\geq \frac{f}{f_0}(\hat{Z}(A), A; \theta) (1 - o_P(1)) \\
 &= \frac{f}{f_0}(\mathcal{Z}, A; \theta) (1 - o_P(1)) \\
 &\quad + \left(\frac{f}{f_0}(\hat{Z}(A), A; \theta) - \frac{f}{f_0}(\mathcal{Z}, A; \theta) \right) (1 - o_P(1)) 1_E \\
 (8) \quad &= \frac{f}{f_0}(\mathcal{Z}, A; \theta) (1 - o_P(1)) + o_P(1)
 \end{aligned}$$

As mentioned in Section 2.3, it holds for all (q, θ) that $\exp[J(q, \theta; A)] \leq g(A; \theta)$. As a result, Theorem 2 implies that

$$\begin{aligned}
 \max_{q \in \mathcal{D}} \frac{\exp[J(q, \theta; A)]}{g(A; \theta_0)} &\leq \frac{g(A; \theta)}{g(A; \theta_0)} \\
 (9) \quad &= \frac{f(A, \mathcal{Z}; \theta)}{f(A, \mathcal{Z}; \theta_0)} (1 + o_P(1)) + o_P(1),
 \end{aligned}$$

where the $o_P(1)$ terms converge uniformly over all $\theta \in \bar{\Theta}$.

Combining Eqs. (8) and (9) yields that for all $\theta \in \bar{\Theta}$, the quantity $\max_{q \in \mathcal{D}} \frac{\exp[J(q, \theta; A)]}{g(A; \theta_0)}$ is upper and lower bounded by $\frac{f}{f_0}(\mathcal{Z}, A; \theta) (1 \pm o_P(1)) + o_P(1)$, establishing the theorem. \square

4. Some statistical applications. With these results, we can show that some standard inference is valid using the likelihood or variational likelihood for blockmodels.

We have that $\hat{\theta}^{VAR} = (\hat{\pi}^{VAR}, \hat{S}^{VAR})$ under P_{θ_0} is asymptotically normal with mean θ_0 and variance covariance matrices given by Theorem 3 and Lemma 2. Since $\theta \mapsto \Sigma(\theta)$ is continuous, we can evidently form tests and confidence regions based on $\sqrt{n}(\hat{\pi}^{VAR} - \pi_0)^T \hat{\Sigma}_1^{-1/2}$ and $\sqrt{n\hat{\lambda}}(\hat{S}^{VAR} - S_0)^T \hat{\Sigma}_2^{-1/2}$, where $\hat{\Sigma}_1$ and $\hat{\Sigma}_2$ are plug-in estimates of Σ_1 and Σ_2 using $\hat{\theta}^{VAR}$, and $\hat{\lambda}$ equals the average degree in the observed data. The same applies to $\hat{\theta}^{ML}$.

Under the CGM standard blockmodel with generative parameter θ_0 , the Wilks statistic is given by

$$\Lambda(\mathcal{Z}, A; \hat{\theta}^{CGM}) \equiv 2 \log \frac{f}{f_0}(\mathcal{Z}, A, \hat{\theta}^{CGM}) \rightarrow \chi^2_{\frac{K(K+3)}{2} - 1}.$$

A consequence of Theorem 2 is that under suitable conditions

$$\Lambda_G(A; \hat{\theta}^{ML}) \equiv 2 \log \frac{g}{g_0}(A; \hat{\theta}^{ML}) = \Lambda(\mathcal{Z}, A; \hat{\theta}^{CGM}) + o_P(1),$$

so that the Wilks statistic for the GM and CGM estimates have the same asymptotic distribution. To see this, observe that Theorem 2 establishes for all $\theta \in \bar{\Theta}$ that

$$\frac{g}{g_0}(A; \theta) = \frac{f}{f_0}(\mathcal{Z}, A; \theta)(1 - \epsilon_n(K, \theta)) + \epsilon_n(K, \theta),$$

where $\sup_{\theta \in \bar{\Theta}} \epsilon_n(K, \theta) = o_P(1)$. By this result, it follows that

$$\begin{aligned} \sup_{\theta \in \bar{\Theta}} \log \frac{g}{g_0}(A; \theta) &= \sup_{\theta \in \bar{\Theta}} \log \left(\frac{f}{f_0}(\mathcal{Z}, A; \theta)(1 - \epsilon_n(K, \theta)) + \epsilon_n(K, \theta) \right) \\ &\leq \sup_{\theta \in \bar{\Theta}} \log \left(\frac{f}{f_0}(\mathcal{Z}, A; \theta)(1 - \epsilon_n(K, \theta)) \right) + o_P(1) \\ &\leq \sup_{\theta \in \bar{\Theta}} \log \frac{f}{f_0}(\mathcal{Z}, A; \theta) + o_P(1) \end{aligned}$$

By similar arguments, it also follows that

$$\sup_{\theta \in \bar{\Theta}} \log \frac{g}{g_0}(A; \theta) \geq \sup_{\theta \in \bar{\Theta}} \log \frac{f}{f_0}(\mathcal{Z}, A; \theta) - o_P(1).$$

Since $\Lambda_G(A; \hat{\theta}^{ML}) = \sup_{\theta \in \bar{\Theta}} \log \frac{g}{g_0}(A; \theta)$ and $\Lambda(\mathcal{Z}, A; \hat{\theta}^{CGM}) = \sup_{\theta \in \bar{\Theta}} \log \frac{f}{f_0}(\mathcal{Z}, A; \theta)$, we have upper and lower bounded $\Lambda_G(A; \hat{\theta}^{ML})$ by $\Lambda(\mathcal{Z}, A; \hat{\theta}^{CGM}) \pm o_P(1)$.

A similar result holds for the Wilks statistic of the variational estimate $\hat{\theta}^{VAR}$. To see this, we observe that since $J(\theta; A) \equiv \max_{q \in \mathcal{D}} \exp[J(q, \theta; A)] \leq g(A; \theta)$, it holds that

$$\frac{J(\theta, A)}{J(\theta_0, A)} \geq \frac{J(\theta; A)}{g(A; \theta_0)},$$

so that Theorem 4 implies

$$\frac{J(\theta, A)}{J(\theta_0, A)} \geq \frac{f}{f_0}(\mathcal{Z}, A; \theta)(1 + o_P(1)) + o_P(1).$$

To upper bound the same quantity, we observe that

$$\begin{aligned} \frac{J(\theta, A)}{J(\theta_0, A)} &\leq \frac{g(A; \theta)}{f(\hat{\mathcal{Z}}, A; \theta_0)} \\ &= \frac{g(A; \theta)}{g(A; \theta_0)f(\hat{\mathcal{Z}}, A; \theta_0)g(A; \theta_0)^{-1}} \\ &= \frac{g(A; \theta)}{g(A; \theta_0)(1 - o_P(1))}, \end{aligned}$$

using similar steps as in the proof of Theorem 4. Thus, the arguments used to bound Λ_G also imply

$$\Lambda_V(\hat{\theta}^{VAR}) \equiv 2 \log \frac{J(\hat{\theta}^{VAR}, A)}{J(\theta_0, A)} = \Lambda(\mathcal{Z}, A; \hat{\theta}^{CGM}) + o_P(1).$$

A third approach to inference, the parametric bootstrap, is also valid for $\hat{\theta}^{VAR}$. The algorithm is

1. Estimate θ by $\hat{\theta}^{VAR}$
2. Generate B graphs of size n according to the blockmodel with parameter $\hat{\theta}^{VAR}$, producing $(\mathcal{Z}_1^*, A_1^*), \dots, (\mathcal{Z}_B^*, A_B^*)$.
3. Fit A_1^*, \dots, A_B^* by variational likelihood to get $\hat{\theta}_1^{VAR*}, \dots, \hat{\theta}_B^{VAR*}$.
4. Compute the variance-covariance matrix of these B vectors and use it as an estimate of the truth, or similarly, use the empirical distribution function of the vectors

THEOREM 5. *Under the conditions of Theorem 2, the parametric bootstrap distribution of $\sqrt{n}(\hat{\pi}^{VAR} - \pi_0)$ and $\sqrt{n\lambda}(\hat{S}^{VAR} - S_0)$ converges to the Gaussian limits given by Lemma 2.*

PROOF. Without loss of generality we take $B = \infty$, so that we are asking that when the underlying parameter is $\hat{\theta}^{VAR}$, the random law of $\sqrt{n}(\hat{\pi}^{VAR*} - \hat{\pi}^{VAR})$ and $\sqrt{n\lambda}(\hat{S}^{VAR*} - \hat{S}^{VAR})$ converges with P_{θ_0} probability tending to 1 to the Gaussian limits of $\sqrt{n}(\hat{\pi}^{CGM} - \pi_0)$ and $\sqrt{n\lambda}(\hat{S}^{CGM} - S_0)$ as generated under θ_0 .

Let $\hat{\pi}^{CGM*}, \hat{S}^{CGM*}$ have the distribution of the CG MLE based on the data that we have generated from $P_{\hat{\theta}^{VAR}}$. By standard exponential theory

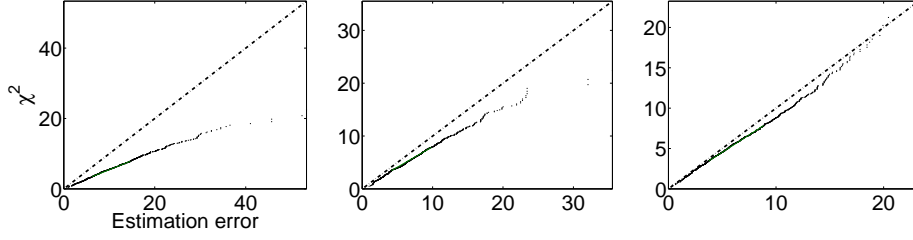


FIG 1. Quantile-quantile plots checking for normality of the estimation error, for synthetically generated blockmodel data with $n = 300, 900$, and 1500 (left to right). The plots suggest convergence to normality as n increases.

such as our Lemma 1, we observe that

$$(10) \quad \sqrt{n}(\hat{\pi}^{CGM*} - \hat{\pi}^{VAR}) \xrightarrow{P_{\hat{\theta}^{VAR}}} N(0, \Sigma_1)$$

$$(11) \quad \sqrt{n\lambda}(\hat{S}^{CGM*} - \hat{S}^{VAR}) \xrightarrow{P_{\hat{\theta}^{VAR}}} N(0, \Sigma_2),$$

since the convergence is uniform on contiguous neighborhoods of θ_0 and the mapping $\theta \rightarrow (\Sigma_1(\theta), \Sigma_2(\theta))$ is smooth. As Theorem 4 implies local asymptotic normality, a theorem of Le Cam [14, Corollary 12.3.1] implies that $P_{\hat{\theta}^{VAR}} \triangleleft P_{\theta_0}$ with P_{θ_0} probability tending to 1, where \triangleleft denotes contiguity. As a result, Le Cam's first contiguity lemma in conjunction with Theorem 4 implies that

$$\begin{aligned} \sqrt{n}(\hat{\pi}^{CGM*} - \hat{\pi}^{VAR*}) &= o_{P_{\hat{\theta}^{VAR}}}(1) \\ \sqrt{n\lambda}(\hat{S}^{CGM*} - \hat{S}^{VAR*}) &= o_{P_{\hat{\theta}^{VAR}}}(1). \end{aligned}$$

Using this result with Eqs. (10), it follows that

$$\begin{aligned} \sqrt{n}(\hat{\pi}^{VAR*} - \hat{\pi}^{VAR}) &\xrightarrow{P_{\hat{\theta}^{VAR}}} N(0, \Sigma_1) \\ \sqrt{n\lambda}(\hat{S}^{VAR*} - \hat{S}^{VAR}) &\xrightarrow{P_{\hat{\theta}^{VAR}}} N(0, \Sigma_2), \end{aligned}$$

establishing the theorem. \square

5. Simulations. Blockmodel parameter estimates π^{VAR}, S^{VAR} were estimated variationally on synthetically generated blockmodel data with $K = 3$ and parameter values $S = \frac{3}{4}(I + 11^T)$, $\pi = [\frac{1}{3} \ \frac{1}{3} \ \frac{1}{3}]$, and $\lambda = \log^2 n$, for a range of graph sizes n . For each value of n , the estimation error $S^{VAR} - S$ from 1000 simulations was recorded.

Figure 1 shows quantile-quantile plots for $n = (300, 900, 1500)$, comparing the empirical distribution of rescaled errors $n\lambda(S^{VAR} - S)^T \Sigma_2^{-1}(S^{VAR} - S)$ to a chi-squared distribution with 6 degrees of freedom. We observe that the distributions grow more similar for large n , as predicted by Lemma 2 and Theorem 3. As an alternative measure of quality, we also compared Z to estimated labels induced from q ; the average fraction of misclassified nodes was .11, .05, and .03

To initialize the variational algorithm, an initial clustering was computed using spectral methods [17], giving an initial guess for S, π and q . The spectral method itself utilized the K-means clustering algorithm, which was initialized randomly. We observed marked improvement by computing multiple initial guesses and using the one for which J was largest.

Appendix: Proof of Theorem 1. Let \hat{Z} be defined to be the maximizer of $\max_{\theta \in \bar{\Theta}} f(\underline{Z}, A; \theta)$ over all $\underline{Z} \in \mathcal{Z}^n$. The following result from [1] then applies to $\hat{Z}(A)$:

THEOREM 6 ([1]). *Let $\rho_n = \omega(n^{-1} \log n)$ and let S have no identical columns. Let $\bar{\Theta} \subset \Theta$ be an open compact set. There exists a sequence $b_n \rightarrow \infty$ such that*

$$\sup_{\theta \in \bar{\Theta}} \mathbb{P}_\theta(Z \neq \hat{Z}(A)) = O(n^{-b_n}).$$

PROOF OF THEOREM 1. We establish that Eqs. (3) and (4) hold. Algebraic manipulation of $\mathbb{P}_\theta(\underline{Z} \neq \hat{Z}(A)|A) \frac{g(A; \theta)}{g(A; \theta_0)}$ produces

$$\mathbb{P}_\theta(\underline{Z} \neq \hat{Z}(A)|A) \frac{g(A; \theta)}{g(A; \theta_0)} = \frac{\mathbb{P}_\theta(A, \underline{Z} \neq \hat{Z}(A))}{g(A; \theta_0)} = \sum_{Z \neq \hat{Z}(A)} \frac{f(Z, A; \theta)}{g(A; \theta_0)}.$$

For fixed θ it holds that

$$\begin{aligned} E_{\theta_0} \left[\mathbb{P}_\theta(\underline{Z} \neq \hat{Z}(A)|A) \frac{g(A; \theta)}{g(A; \theta_0)} \right] &= E_{\theta_0} \left[\frac{\mathbb{P}_\theta(A, \underline{Z} \neq \hat{Z}(A))}{g(A; \theta_0)} \right] \\ &= \sum_A \frac{\mathbb{P}_\theta(A, \underline{Z} \neq \hat{Z}(A))}{g(A; \theta_0)} \mathbb{P}_{\theta_0}(A) \\ &= P_\theta(Z \neq \hat{Z}(A)). \end{aligned}$$

Markov's inequality and Theorem 6 imply that for any fixed θ and fixed $\epsilon > 0$,

$$\mathbb{P}_{\theta_0} \left\{ \mathbb{P}_\theta(\underline{Z} \neq \hat{Z}(A)|A) \frac{g(A; \theta)}{g(A; \theta_0)} \geq \epsilon \right\} \leq \frac{\mathcal{O}(n^{-b_n})}{\epsilon}.$$

To show uniform convergence over all $\theta \in \bar{\Theta}$, let $\bar{\Theta}^{(n)}$ denote a set of points which induce an n^{-3} covering of $\bar{\Theta}$ in the sup norm. Let $\theta^{(i)} \equiv (\pi^{(i)}, H^{(i)})$ denote the i th point in $\bar{\Theta}^{(n)}$. It holds for any $\theta' \equiv (\pi', H')$ in the ball $B(\theta^{(i)}, n^{-3})$ and for any Z that

$$\begin{aligned}
\frac{f(Z, A; \theta')}{f(Z, A; \theta^{(i)})} &= \left(\prod_{i=1}^n \frac{\pi'(Z_i)}{\pi^{(i)}(Z_i)} \right) \left(\prod_{i>j} \left(\frac{H'(Z_i, Z_j)}{H^{(i)}(Z_i, Z_j)} \right)^{A_{ij}} \left(\frac{1 - H'(Z_i, Z_j)}{1 - H^{(i)}(Z_i, Z_j)} \right)^{1-A_{ij}} \right) \\
&\leq \left(\prod_{i=1}^n 1 + \frac{1}{n^3 \pi^{(i)}(Z_i)} \right) \left(\prod_{i>j} \left(1 + \frac{1}{n^3 H^{(i)}(Z_i, Z_j)} \right)^{A_{ij}} \right. \\
&\quad \cdot \left. \left(1 + \frac{1}{n^3 (1 - H^{(i)}(Z_i, Z_j))} \right)^{1-A_{ij}} \right) \\
&\leq \left(1 + C_{\mathcal{M}} \frac{1}{n^2} \right)^{n+n(n+1)/2} \\
&= \mathcal{O}(1),
\end{aligned}$$

where $C_{\mathcal{M}}$ is a constant which depends on $H^{(i)}$ being strictly bounded away from $n^{-1} \log n$ or 1 in each coordinate which is allowed to vary in Φ ; because Φ is open and compact and $\rho > n^{-1} \log n$, this will always hold.

It follows that for $\theta' \in B(\theta^{(i)}, 1/n^3)$,

$$\begin{aligned}
\mathbb{P}_{\theta'}(\mathcal{Z} \neq \hat{Z}(A) | A) \frac{g(A; \theta')}{g(A; \theta_0)} &= \sum_{Z \neq \hat{Z}(A)} \frac{f(Z, A; \theta')}{g(A; \theta_0)} \\
&= \sum_{Z \neq \hat{Z}(A)} \frac{f(Z, A; \theta^{(i)})}{f(Z, A; \theta^{(i)})} \frac{f(Z, A; \theta')}{g(A; \theta_0)} \\
&= \sum_{Z \neq \hat{Z}(A)} \mathcal{O}(1) \frac{f(Z, A; \theta^{(i)})}{g(A; \theta_0)} \\
&= \mathcal{O}(1) \mathbb{P}_{\theta^{(i)}}(\mathcal{Z} \neq \hat{Z}(A) | A) \frac{g(A; \theta^{(i)})}{g(A; \theta_0)}
\end{aligned}$$

and hence

$$\begin{aligned} & \mathbb{P}_{\theta_0} \left\{ \sup_{\theta \in \bar{\Theta}} \mathbb{P}_{\theta}(\mathcal{Z} \neq \hat{Z}(A)|A) \frac{g(A; \theta)}{g(A; \theta_0)} \geq \epsilon \right\} \\ & \leq \sum_{i=1}^{|M^{(n)}|} \mathbb{P}_{\theta_0} \left\{ \sup_{\theta \in B(\theta^{(i)}, \epsilon/n^2)} \mathbb{P}_{\theta}(\mathcal{Z} \neq \hat{Z}(A)|A) \frac{g(A; \theta)}{g(A; \theta_0)} \geq \epsilon \right\} \\ & \leq \mathcal{O}(1) \cdot |\bar{\Theta}^{(n)}| \cdot \mathcal{O}(n^{-b_n})/\epsilon. \end{aligned}$$

Substituting $|\bar{\Theta}^{(n)}| = \mathcal{O}(n^{3(K+1)K/2+3(K-1)})$ and recalling that $\gamma_n \rightarrow \infty$ completes the proof, as the right hand side converges to zero for any $\epsilon > 0$. This establishes Eq. (3). The remaining equation, Eq. (4), is given by the result of Theorem 6. \square

References.

- [1] BICKEL, P. J. and CHEN, A. (2009). A nonparametric view of network models and Newman–Girvan and other modularities. *Proc. Natl. Acad. Sci. USA* **106** 21068.
- [2] BICKEL, P. J., CHEN, A. and LEVINA, E. (2011). The method of moments and degree distributions for network models. *Ann. Statist.* **39** 38–59.
- [3] BICKEL, P. and DOKSUM, K. (1977). *Mathematical Statistics: Basic ideas and selected topics*. Holden-Day, San Francisco.
- [4] BOLLOBÁS, B., JANSON, S. and RIORDAN, O. (2007). The phase transition in inhomogeneous random graphs. *Random Structures Algorithms* **31** 3–122.
- [5] CELISSE, A., DAUDIN, J. J. and PIERRE, L. (2011). Consistency of maximum-likelihood and variational estimators in the Stochastic Block Model. *Arxiv preprint arXiv:1105.3288*.
- [6] CHANNAROND, A., DAUDIN, J. J. and ROBIN, S. (2011). Classification and estimation in the Stochastic Block Model based on the empirical degrees. *Arxiv preprint arXiv:1110.6517*.
- [7] CHOI, D. S., WOLFE, P. J. and AIROLDI, E. M. (2012). Stochastic blockmodels with growing number of classes. *Biometrika* **99** 273–284.
- [8] COJA-OGHLAN, A. and LANKA, A. (2008). Partitioning Random Graphs with General Degree Distributions. In *Fifth IFIP International Conference On Theoretical Computer Science–TCS 2008* 127–141.
- [9] DAUDIN, J. J., PICARD, F. and ROBIN, S. (2008). A mixture model for random graphs. *Statist. Comput.* **18** 173–183.
- [10] HANDCOCK, M. S., RAFTERY, A. E. and TANTRUM, J. M. (2007). Model-based clustering for social networks. *J. Roy. Statist. Soc. Ser. A* **170** 301–354.
- [11] HOFF, P. D., RAFTERY, A. E. and HANDCOCK, M. S. (2002). Latent space approaches to social network analysis. *J. Am. Statist. Assoc.* **97** 1090–1098.
- [12] KARRER, B. and NEWMAN, M. (2011). Stochastic blockmodels and community structure in networks. *Phys. Rev. E* **83** 016107.
- [13] LAZER, D., PENTLAND, A. S., ADAMIC, L., ARAL, S., BARABASI, A. L., BREWER, D., CHRISTAKIS, N., CONTRACTOR, N., FOWLER, J., GUTMANN, M. et al. (2009). Life in the network: the coming age of computational social science. *Science* **323** 721.

- [14] LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing statistical hypotheses*. Springer Verlag.
- [15] PROULX, S. R., PROMISLOW, D. E. L. and PHILLIPS, P. C. (2005). Network thinking in ecology and evolution. *Trends in Ecology & Evolution* **20** 345–353.
- [16] ROHE, K., CHATTERJEE, S. and YU, B. (2011). Spectral clustering and the high-dimensional stochastic blockmodel. *Ann. Statist.* **39** 1878–1915.
- [17] VON LUXBURG, U. (2007). A tutorial on spectral clustering. *Statist. Comput.* **17** 395–416.

E-MAIL: bickel@stat.berkeley.edu

E-MAIL: dchoi@stat.berkeley.edu

E-MAIL: xiangyuchang@gmail.com

E-MAIL: zhanghai@nwu.edu.cn